Application of Expressway Big Data in Highway Freight Traffic Statistics

Department of Service Statistics, NBS

Oct 28. 2014

Part A Research background

Part B Expressway Big Data and preprocessing

Part C Analysis of overloaded vehicles behaviors based on Big Data

Part D Inspiration and prospect

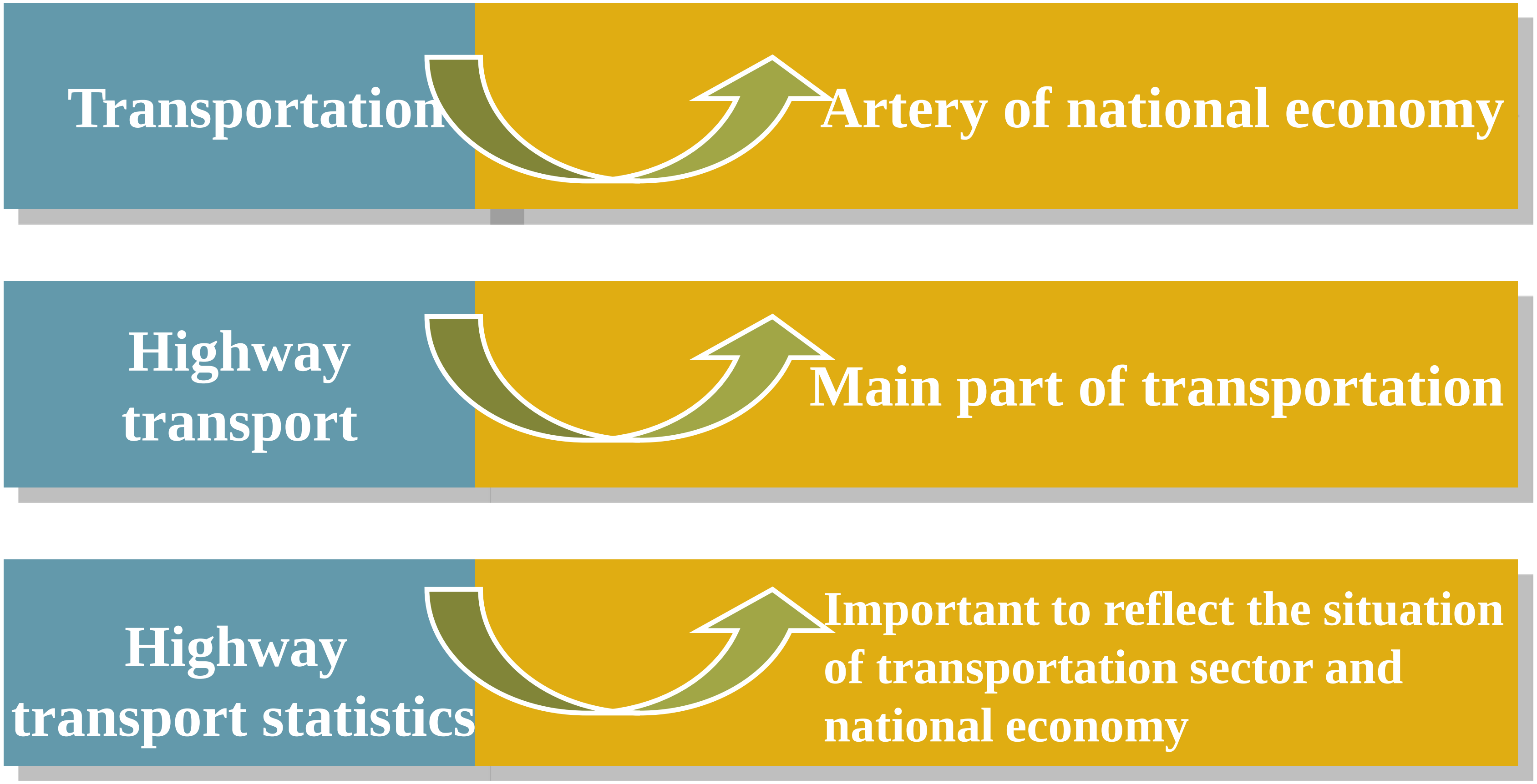
Part A Research background

- Importance and difficulties of highway transport statistics
- New highway transport statistics methods based on Big Data

PART A BACKGROUND

Importance and difficulties of highway transport statistics

Importance



Transportation → Artery of national economy

Highway transport → Main part of transportation

Highway transport statistics → Important to reflect the situation of transportation sector and national economy

Difficulties

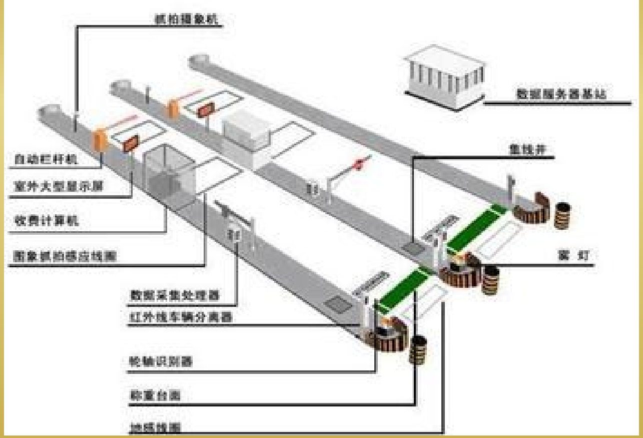
Restricted by the industry’s own characteristics, there are some difficulties in highway transport statistics mainly caused by vehicle mobility and instability of operators. Especially in the highway freight traffic sector whose market access threshold is low, the market is full of small size operators and their number is always very large. And also, these operators always move, collapse or suddenly come out many. All these bring difficulties to the highway transport statistics and reduce its efficiency.

PART A

BACKGROUND

New highway transport statistics methods based on Big Data

Department of Service Statistics combined with Department of Comprehensive Planning & Design, Ministry of Transport developed the new highway transport statistics methods based on Big Data: get the monthly traffic base by the traditional sampling method and calculate the monthly fluctuation coefficient through Big data. Monthly traffic base multiplied by fluctuation coefficient gives monthly traffic.



Fluctuation coefficient of passenger traffic
——Passenger station ticket record

Fluctuation coefficient of freight traffic
——Weight toll system record

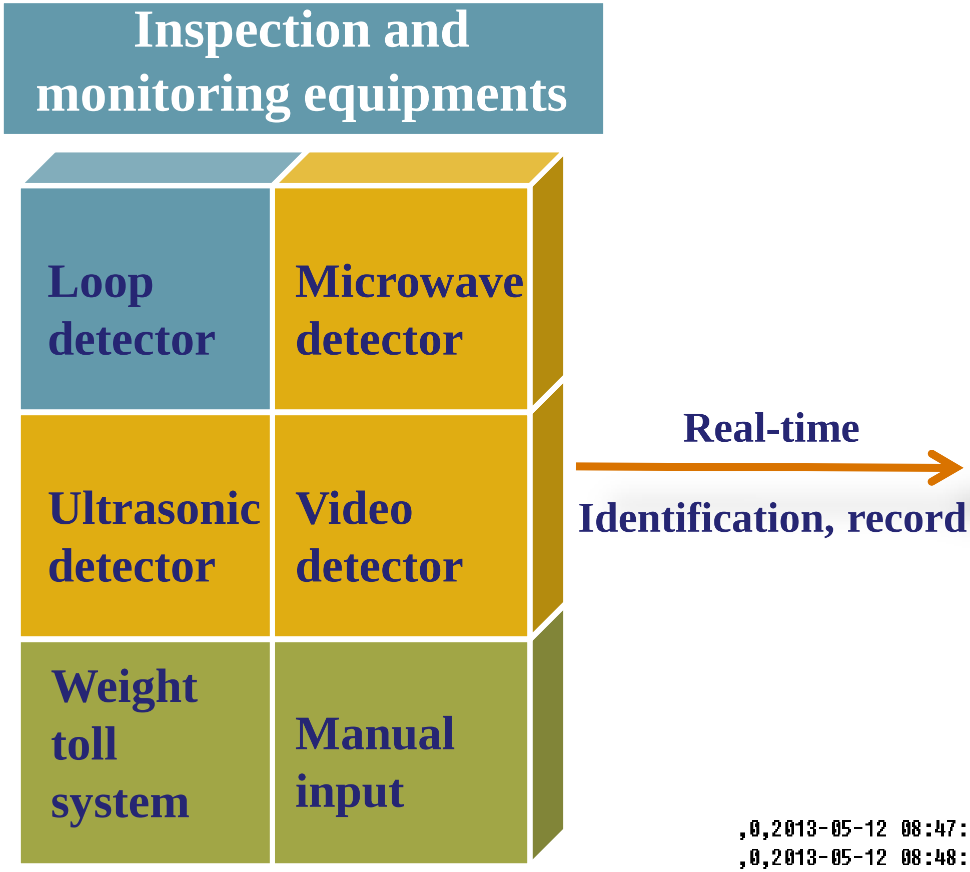
There are still problems in the new method, so Department of Service statistics made further study about the expressway Big Data, hoping to further improve the highway freight traffic statistics methods and explore ways to combine traditional statistics with Big Data.

Part B Expressway Big Data and preprocessing

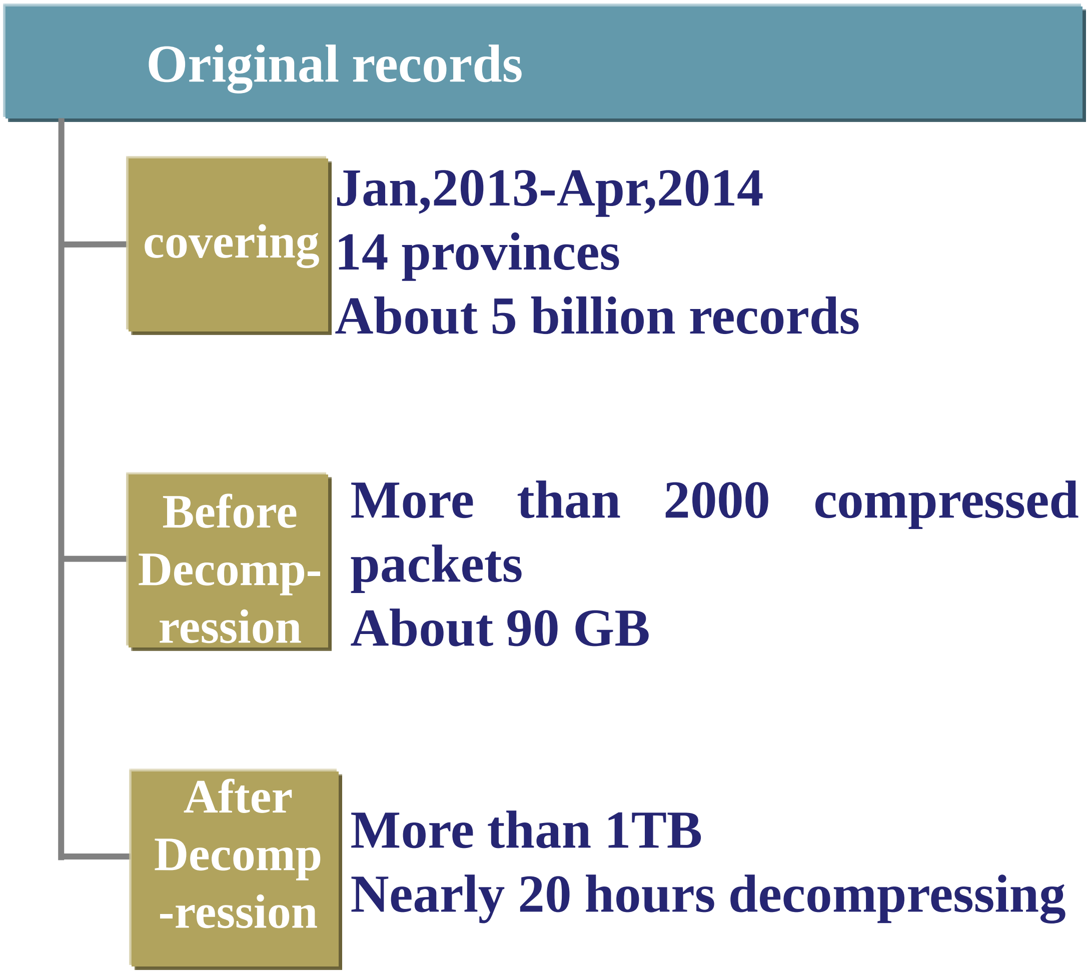
- Big Data sources
- Preprocessing

Big Data sources

Original records of expressway monitoring system



```
,0,2013-05-12 08:47:55.0,,2301,2013-05-12 08:47:55.0,蓝皖ASM113,1,0,0,0,,0,0
,0,2013-05-12 08:48:14.0,,2301,2013-05-12 08:48:14.0,蓝苏A9NP86,1,0,0,0,,0,0
,0,2013-05-12 08:48:32.0,,2301,2013-05-12 08:48:32.0,蓝皖AZ6571,1,0,0,0,,0,0
,0,2013-05-12 08:48:50.0,,2301,2013-05-12 08:48:50.0,蓝皖AZ5812,1,0,0,0,,0,0
```



Original records

covering

Jan,2013-Apr,2014
14 provinces
About 5 billion records

Before Decomp-ression

More than 2000 compressed packets
About 90 GB

After Decomp-ression

More than 1TB
Nearly 20 hours decompressing

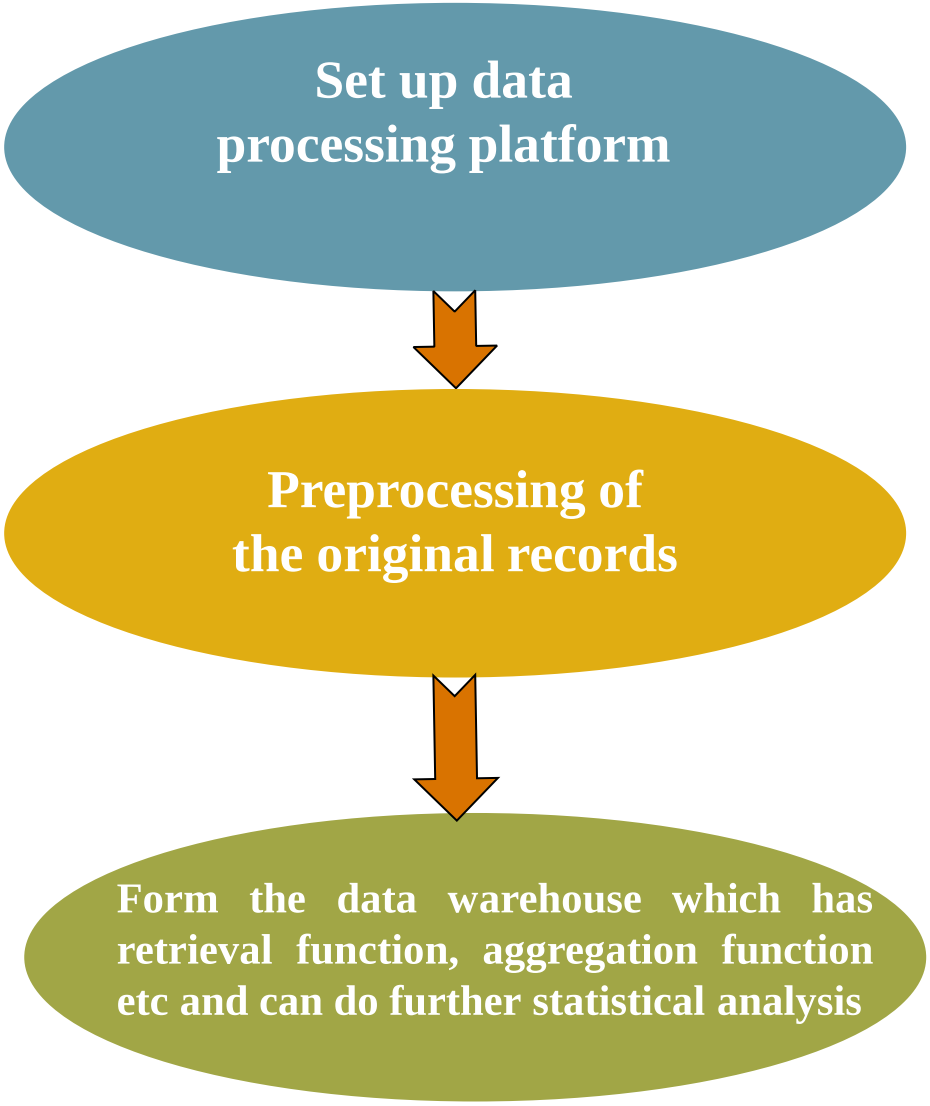
Consitent with ‘4V’ characteristics

- Volume
- Velocity
- Variety
- Value

PART B

SOURCE&PREPROCESSING

Preprocessing

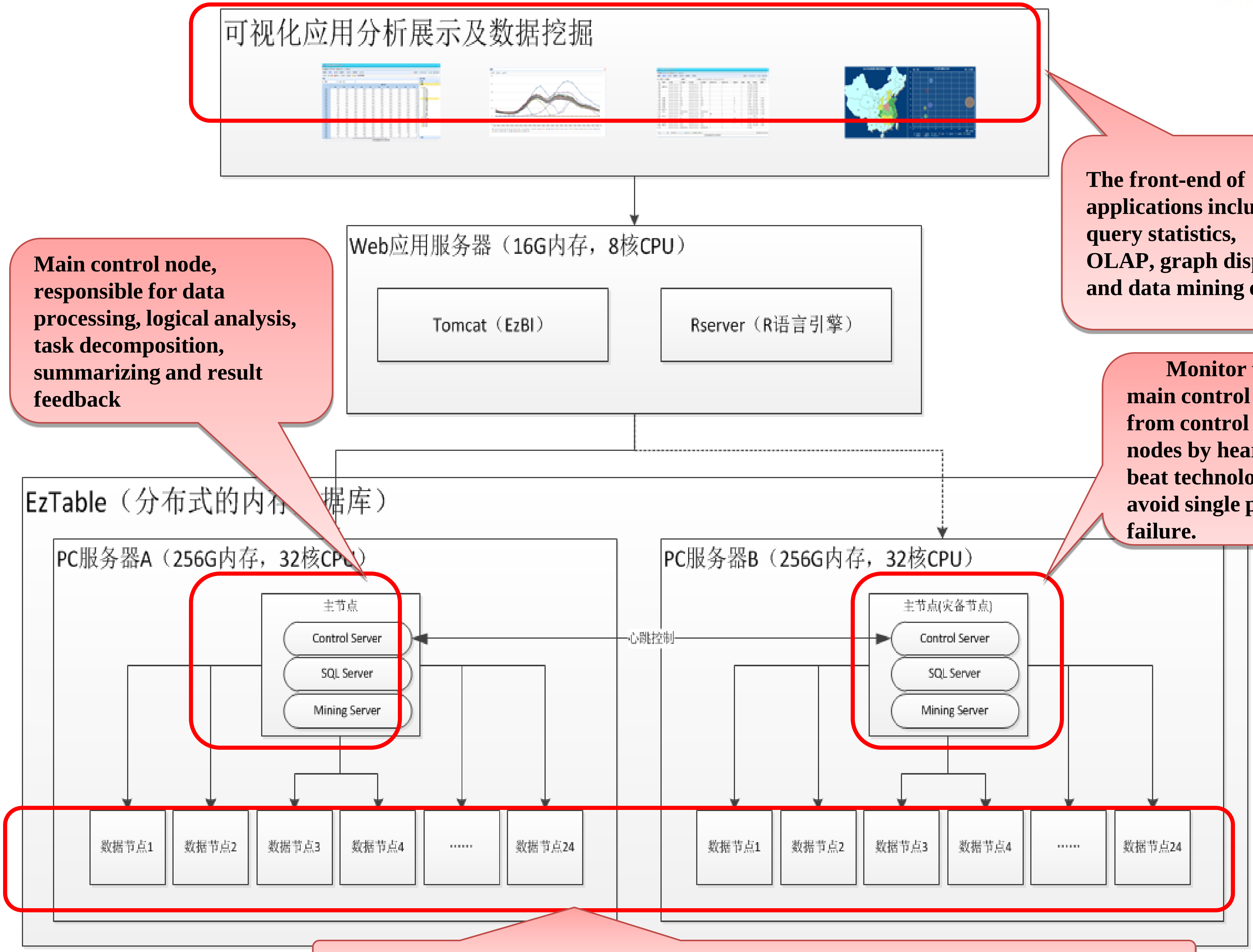


Set up data processing platform

Preprocessing of the original records

Form the data warehouse which has retrieval function, aggregation function etc and can do further statistical analysis

Set up data preprocessing platform



The front-end of applications include query statistics, OLAP, graph display and data mining etc.

Main control node, responsible for data processing, logical analysis, task decomposition, summarizing and result feedback

Monitor the main control node from control nodes by heart-beat technology to avoid single point failure.

Two PC servers with 24 nodes each

PART B

SOURCE&PREPROCESSING

Preprocessing of the original records

There are differences among the provinces data in format, meaning and other fields.

What we did?

Removed or adjusted some invalid data, abnormal data.

Added some incomplete data which can be estimated.

Unified the format, meaning and code.

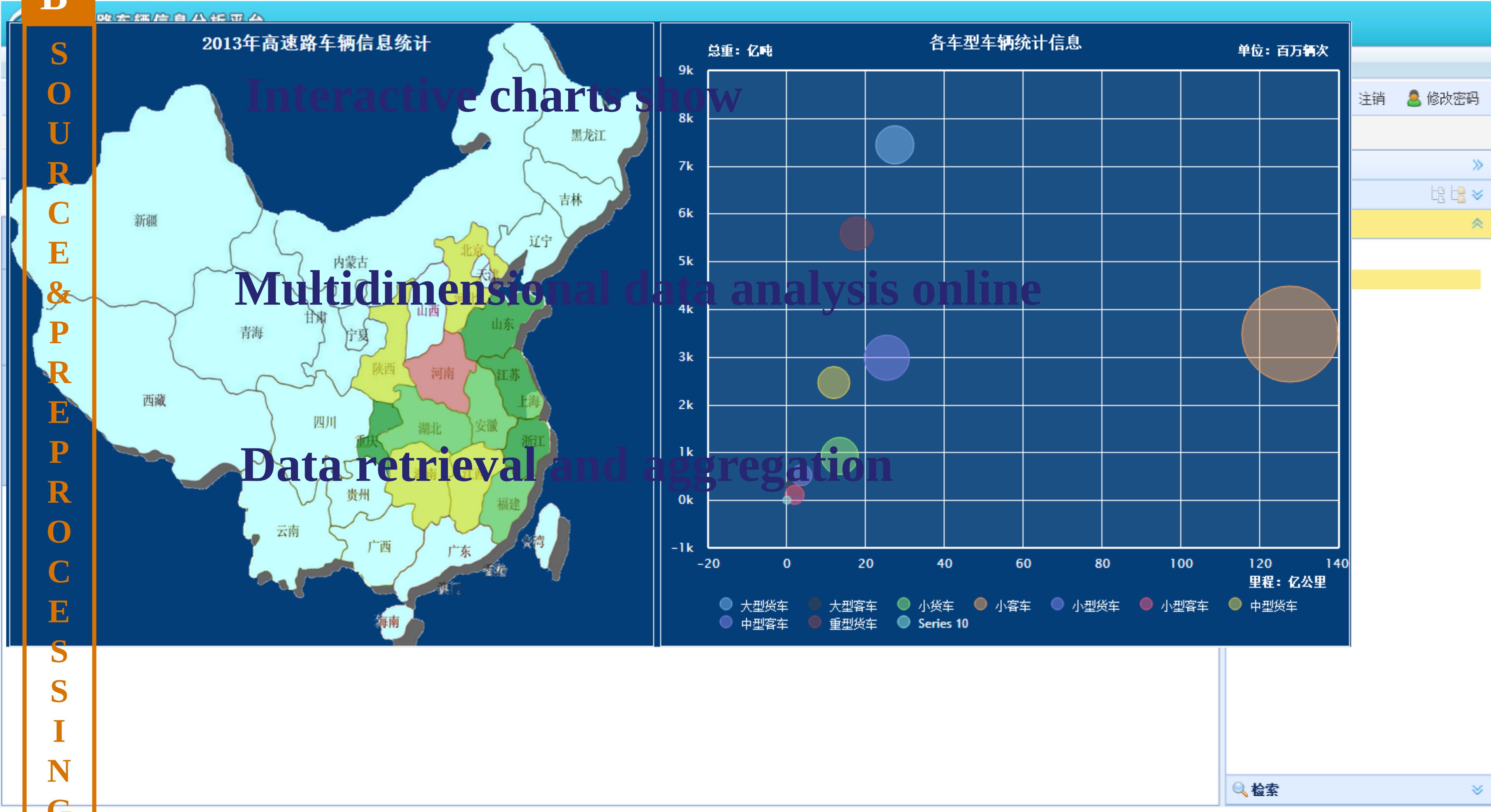
Example: Vehicle type code

Province	example	illustration
anhui	1-4，11-15	Standard code
fujian	1-4，1-5	1-5 convert to 11-15
guangdong		Convert according to rule
hebei	1-4，1-9	1-5 convert to11-15，6-9 convert to 15
hbjspq	1-4，11-15	Standard code
henan		Convert according to rule
hubei		Convert according to rule
hunan	1-4，1-9	1-4 convert to 11-14，5-9 convert to 15
jiangsu	1-4，11-15	Standard code
jiangxi	1-4，11-22	21-22 convert to 15
shandong	1-4，11-15	Standard code
shaanxi	1-4，1-5	1-5 convert to 11-15
shanghai	1-4，5-11	5-9 convert to 11-15，10-11convert to 15
zhejiang	1-4，1-7	1-5convert to11-15，6-7convert to15
chongqing	1-4，1-5	1-5 convert to 11-15

PART B

SOURCE & PREPROCESSING

Form the data warehouse which has retrieval function, aggregation function etc and can do further statistical analysis



Interactive charts show

Multidimensional data analysis online

Data retrieval and aggregation

Part C Analysis of overloaded vehicles behaviors based on Big Data

- Goals
- Ideas and methods
- Implementation process
- Preliminary conclusions

PART C

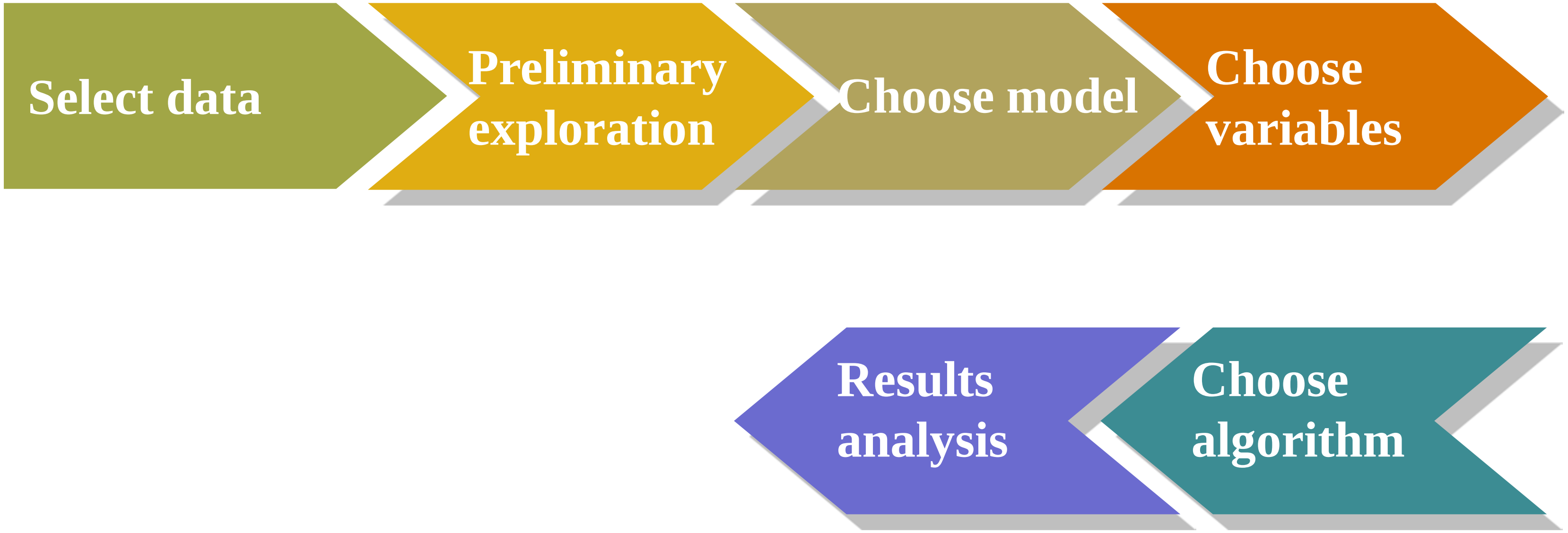
OVERLOADED VEHICLES

Goals

1. Get familiar with mining methods in Big Data
2. Know more about variables
3. Lay the foundation for the improvement of highway freight traffic statistics method
4. Gather experience for Big Data application

Ideas and methods

Do preliminary mining in Big Data to find the characteristics of the variables. Establish reasonable models based on these characteristics and choose feasible algorithm to do deep mining. Always change during the process and analyze the model results.



Select data

Preliminary exploration

Choose model

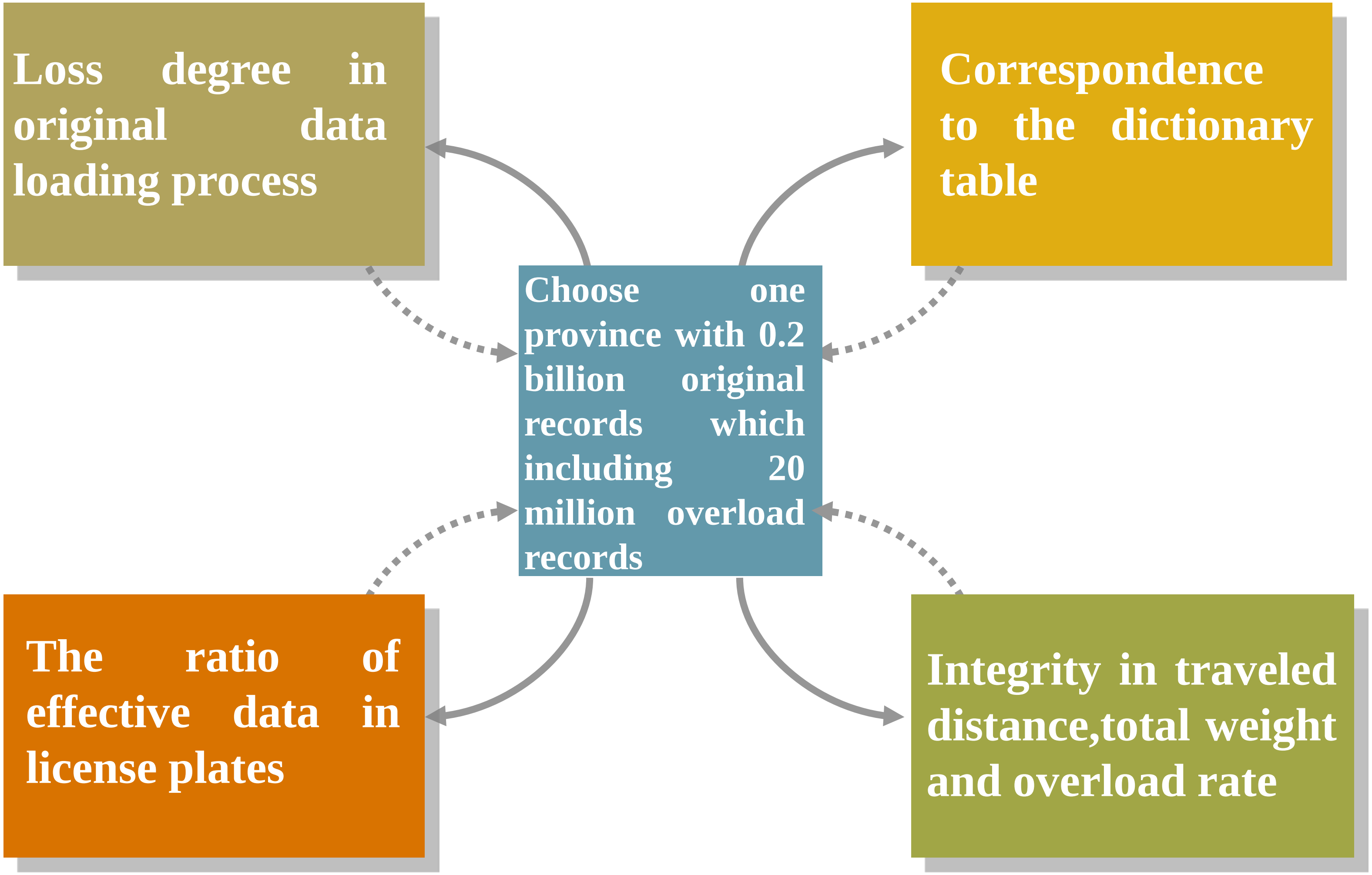
Choose variables

Choose algorithm

Results analysis

Implementation process

Select data： choose one province with good data quality



Choose one province with 0.2 billion original records which including 20 million overload records

Loss degree in original data loading process

Correspondence to the dictionary table

The ratio of effective data in license plates

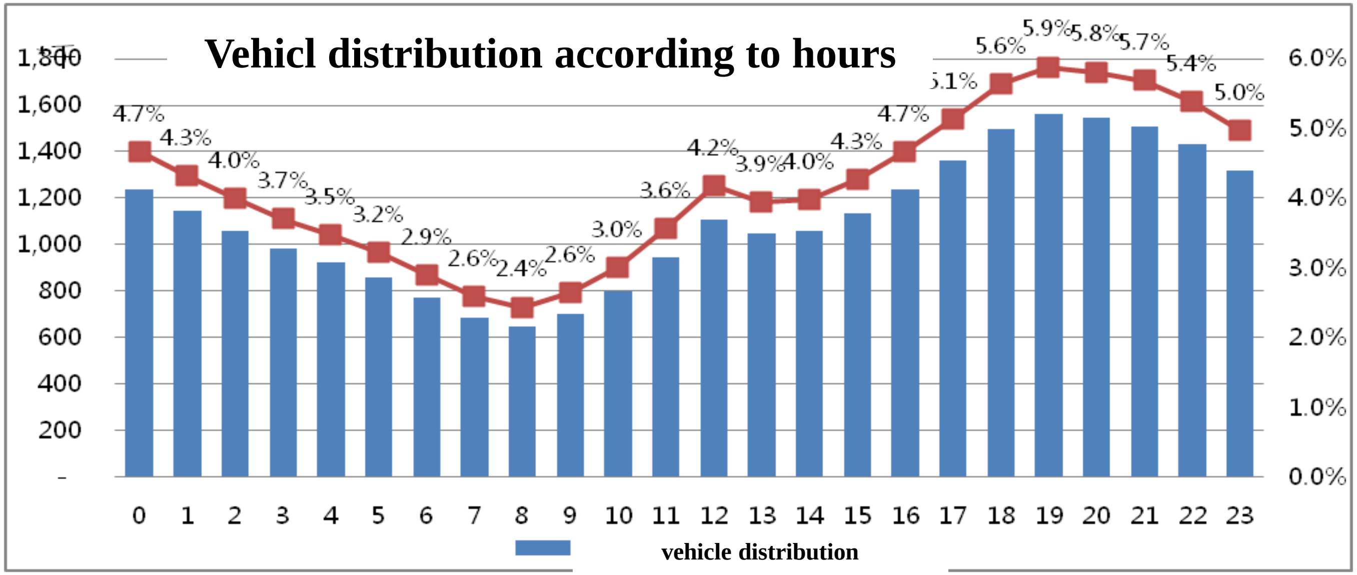
Integrity in traveled distance,total weight and overload rate

PART C

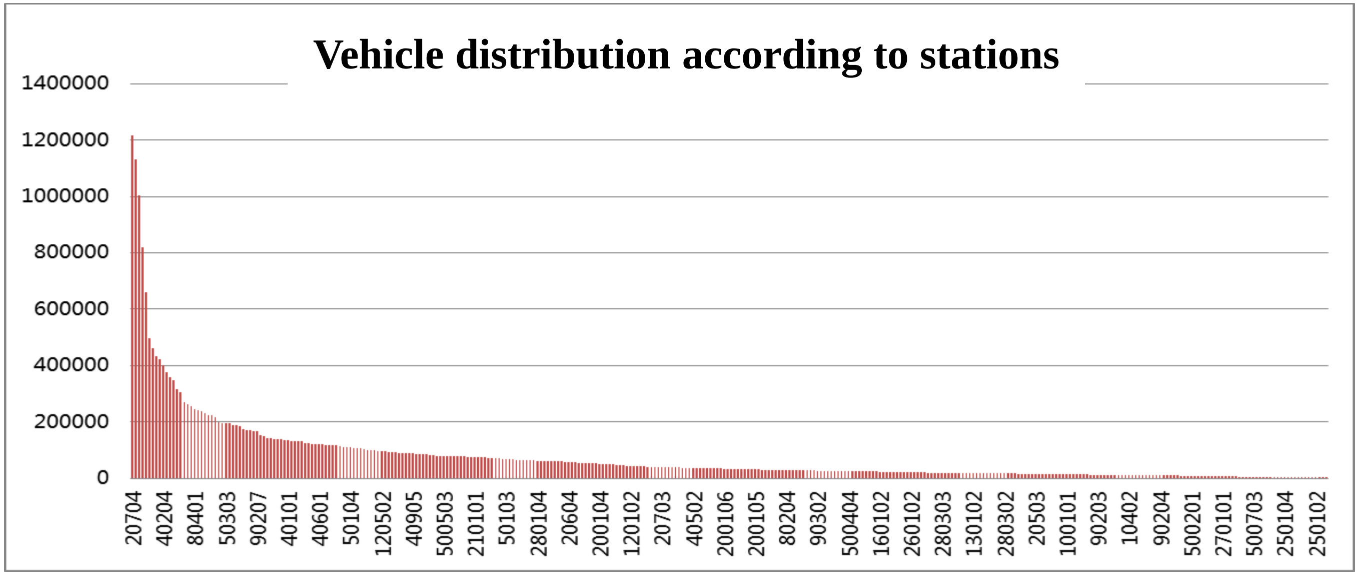
OVERLOADED VEHICLES

Preliminary exploration of variables

Hours



Entrances

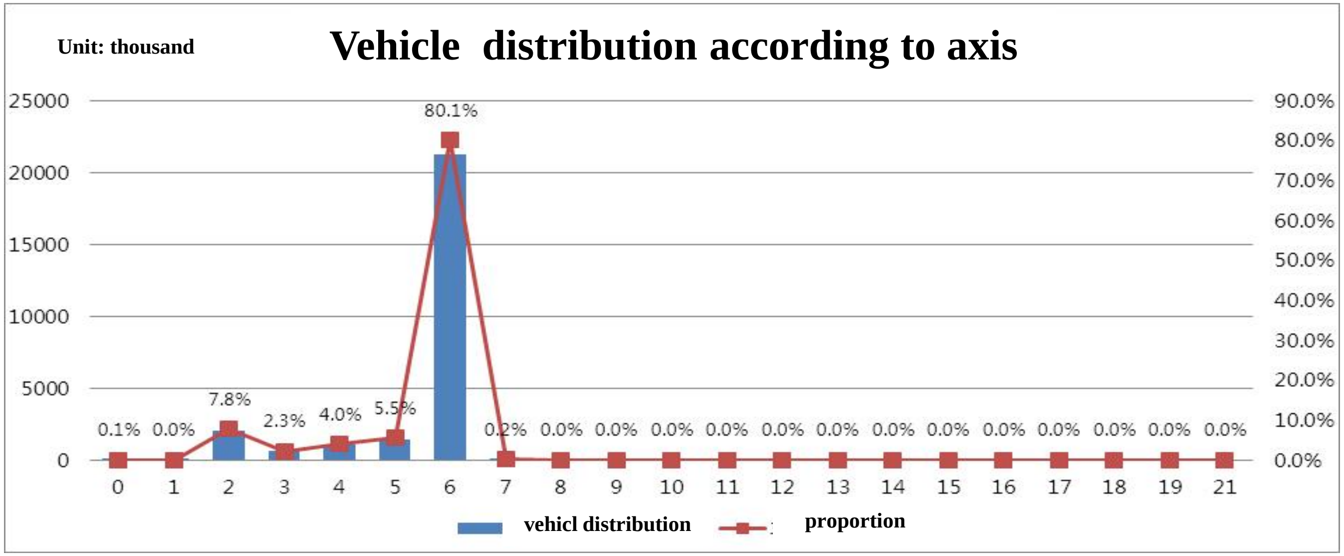


PART C

OVERLOADED VEHICLES

Axis

DiStance,weight



quantile	distance	total weight	count
0%	-	-	1
5%	15	18,100	1
10%	25	22,400	1
15%	39	29,600	1
20%	54	36,600	1
25%	75	39,200	1
30%	100	45,200	1
35%	130	52,500	2
40%	166	57,000	2
45%	210	73,600	2
50%	263	87,300	3
55%	324	105,300	3
60%	395	127,100	4
65%	491	161,500	5
70%	640	211,100	7
75%	858	290,400	9
80%	1,239	432,700	12
85%	2,003	729,100	19
90%	4,015	1,562,600	36
95%	11,705	5,030,000	96
100%	1,380,436	331,668,700	6,336

PART C

OVERLOADED VEHICLES

Choose model and variables

According to the characteristics of variables found through preliminary exploration and also considering the data size, we tried several models, compared their feasibility and results. Finally, we chose clustering model.

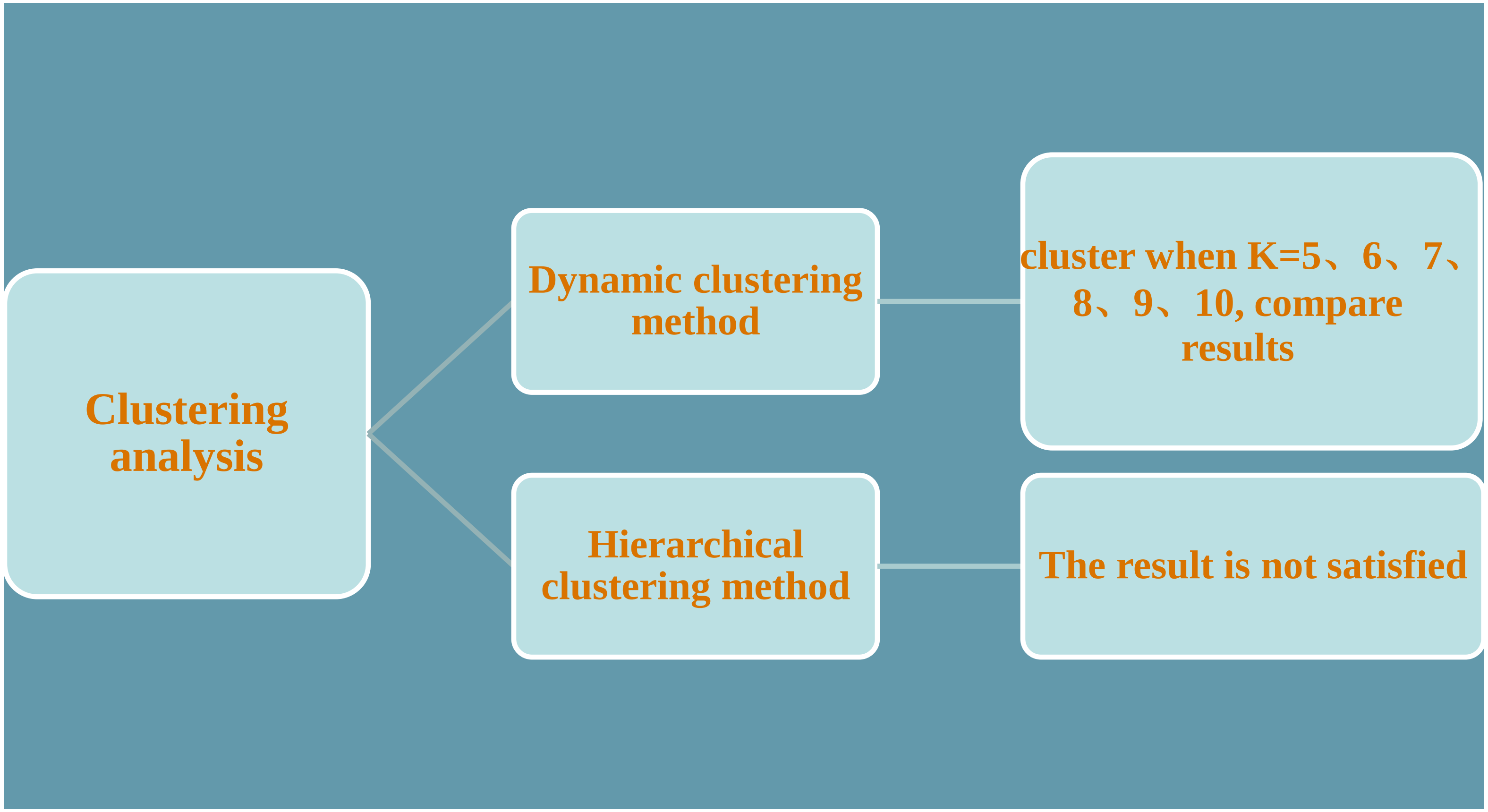
Select variables based on their explanation and importance to overload behaviors. The variables put in model are: month, hour, distance, weight and count of vehicles.

variable code	varible name	illustration	in or not in model
V1	PROVINCE	only one province	no
V2	YEAR	year 2013 and 2014	no
V3	ENTRY_mon	month to enter expressway	yes
V4	ENTRY_hour	hour to enter expressway 0-23	yes
V5	ENTRY_station	identify number of entrance station	no
V6	VEHICLE_class_s	class of vehicle	no
V7	VEHICLE_type_S	type of vehicle	no
V8	AXIS_num	axis number of vehicle	no
V9	sum(DISTANCE）	distance traveled on expressway	yes
V10	sum(TOTAL_weight）	the total waight of vehicle and freight	yes
V11	COUNT	count of vehicle	yes

PART C

OVERLOADED VEHICLES

Choose algorithm



Clustering analysis

Dynamic clustering method

cluster when K=5、6、7、8、9、10, compare results

Hierarchical clustering method

The result is not satisfied

Results analysis

When k=8, the result is better than when k=5,6,7,9,10, so we do clustering analysis in 8 layers.

varible	group 1	group 2	group 3	group 4	group 5	group 6	group 7	group 8
ENTRY_mon	6	7	2	3	10	10	3	6
ENTRY_hour	13	12	12	4	5	18	20	13
sum(DISTANCE)	586543	1616	1646	1489	1542	2044	2183	99034
sum(TOTAL_weight)	138981	618	578	704	728	711	738	38653
COUNT	2632	13	12	14	15	15	16	685
NUMBER of PARTICLE DATA	1483	119328	170986	179667	157086	182685	205430	11625
PROPORTION	0.14%	11.60%	16.63%	17.47%	15.28%	17.77%	19.98%	1.13%

PART C

OVERLOADED VEHICLES

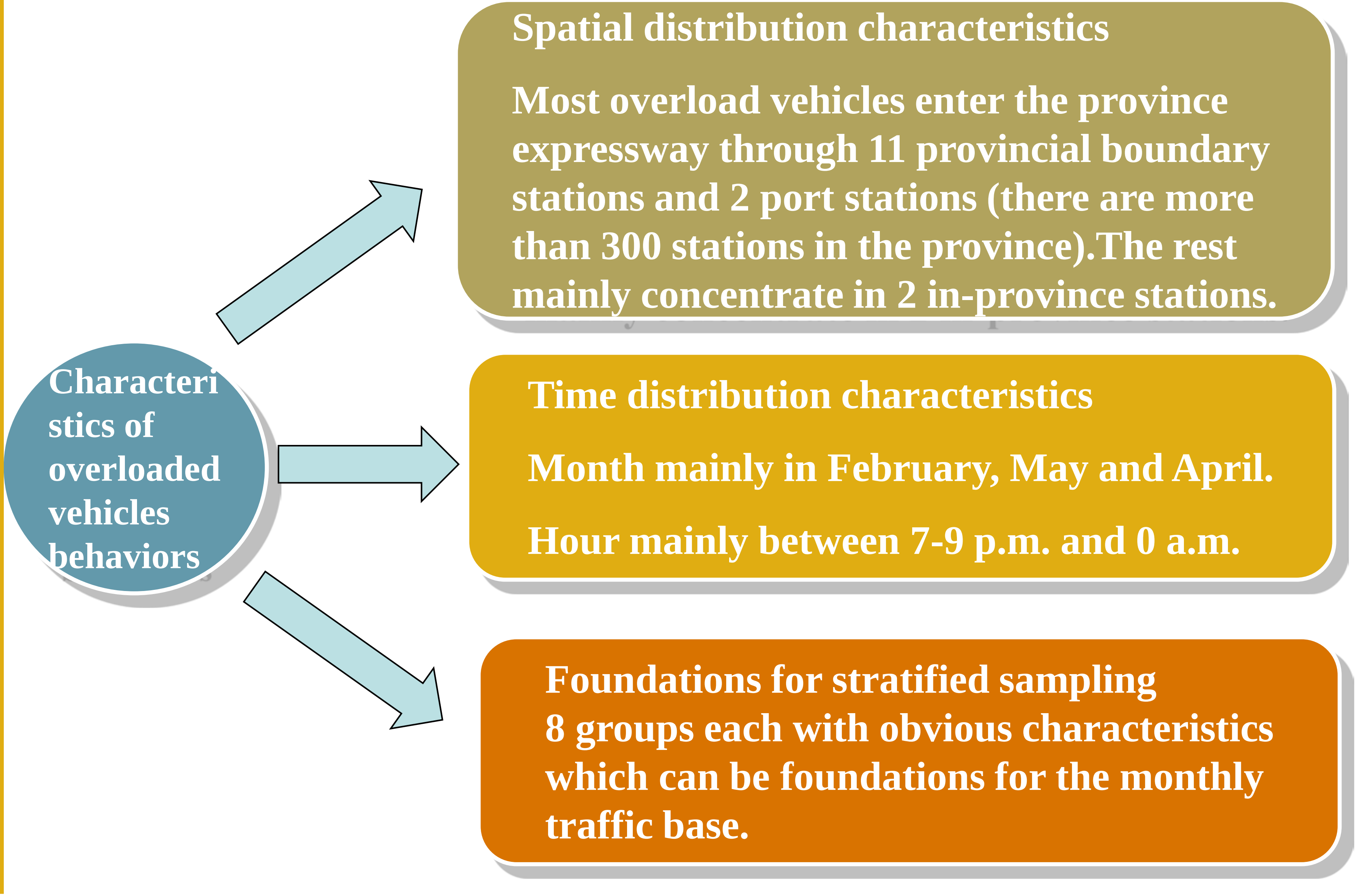
	group 1	group 2	group 3	group 4	group 5	group 6	group 7	group 8
total distance	869842656	192803531	281469518	267586532	242202921	373384933	448405833	1151266010
proportion	22.73%	5.04%	7.35%	6.99%	6.33%	9.76%	11.72%	30.08%
average distance	222	121	132	103	103	133	138	144
total weight	206109379067	73690664829	98857784998	126413106809	114330606021	129899985706	151697784241	449339267947
proportion	15.26%	5.46%	7.32%	9.36%	8.47%	9.62%	11.23%	33.28%
total count	3902859	1587032	2125577	2590700	2350762	2800574	3246403	7964479
proportion	14.69%	5.97%	8.00%	9.75%	8.85%	10.54%	12.22%	29.98%

group 1	Station no	120106	20101	20704	50305	20801	50304	60108	230207	50205	240104
	number of particle data	334	324	306	123	105	64	62	54	48	30
	location	boundary	boundary	boundary	boundary	boundary	boundary	boundary	boundary	in-province	boundary
group 8	Station no	20505	40204	120201	50304	230207	240104	80109	60108	503010	50305
	number of particle data	364	348	327	310	306	303	297	284	260	259
	location	boundary	boundary	in-province	boundary	boundary	boundary	port	boundary	port	boundary

PART C

OVERLOADED VEHICLES

Preliminary conclusion



Characteristics of overloaded vehicles behaviors

Spatial distribution characteristics

Most overload vehicles enter the province expressway through 11 provincial boundary stations and 2 port stations (there are more than 300 stations in the province).The rest mainly concentrate in 2 in-province stations.

Time distribution characteristics

Month mainly in February, May and April.

Hour mainly between 7-9 p.m. and 0 a.m.

Foundations for stratified sampling

8 groups each with obvious characteristics which can be foundations for the monthly traffic base.

Part D Inspiration and prospect

- » Inspiration of Big Data application
- » Prospect of the combination between Big Data and transport statistics

PART D

INSPIRATION&PROSPECT

Inspiration of Big Data application

The directions of analysis could only be found through a lot of exploration and are always in change.

Simple algorithms are often more practical and effective than complex algorithms.

Getting the data, preprocessing and establishing data warehouse are the most important parts.

PART D

INSPIRATION&PROSPECT

Prospect of the combination of Big Data and transport statistics

Continue the analysis of overloaded vehicles behaviors and consider their effect when calculating the monthly traffic base.

Find further relationship between expressway traffic and highway traffic in order to improve the method of calculating fluctuation coefficient.

Get more data and study the relationship between expressway traffic and national economy.

Thanks！